Identification of Bibliographic Information Written in both Japanese and English

Yuko Taniguchi
Toshiba Solutions Corporation
Taniguchi.Yuko@toshiba-sol.co.jp

Hidetsugu Nanba
Hiroshima City University
nanba@hiroshima-cu.ac.jp

Abstract. We have studied the automatic construction of a multilingual citation index by collecting Postscript and PDF files from the Internet. We propose a method to identify duplicate bibliographic information written in both Japanese and English, which will be an indispensable module for the construction of a multilingual citation index. To confirm the effectiveness of our method, we conducted an examination and found that our method obtained a precision of 94.8% and a recall of 25.6%.

Key words: translation of technical terms, multilingual citation index, identification of interlingual bibliographic information.

1 Introduction

We have studied the automatic construction of a multilingual citation index by collecting Postscript and PDF files from the Internet [2], and in this paper, we propose a method that can identify duplicate bibliographic information written in both Japanese and English, which will be an indispensable module for the construction of a multilingual citation index.

There are several related citation indices, such as Google Scholar (http://www.google.com/schhp), CiteSeer (http://citeseer.ist.psu.edu/), and PRESRI (http://www.presri.com) [2], which include Postscript and PDF files located on the World Wide Web (WWW). These indices are constructed in two stages: (1) the extraction and (2) the integration of bibliographic information. In the extraction stage, bibliographic information and a list of references are extracted from each paper (Postscript and PDF), and in the integration stage, the extracted bibliographic information is gathered and integrated. A key technique in this stage is to identify any duplicate bibliographic information, and methods of identifying intralingual bibliographic information have already been proposed and implemented in previous systems. However, interlingual identification of bibliographic information is required to construct a multilingual citation index. The following text is an example of the same bibliographic information written in Japanese and English.

[Japanese title]
山田太郎, サポートベクトルマシンを用いた自動要約, Vol. 30, No. 18, pp. 10-21, 2000.
[English title]
Taro Yamada, (2000) "Automatic Summarization based on Support Vector Machines," Journal of Natural Language Processing, 30 (18), pages 10-21. (in Japanese)

As shown in this example, “in Japanese” is written after the bibliographic information in a list of references when a Japanese language paper is cited in an English language paper. In such cases, the corresponding Japanese bibliographic information should be detected. However, traditional systems cannot identify this bibliographic information and therefore we focused on the identification of duplicate interlingual bibliographic information.

2 Identification of Interlingual Bibliographic Information

2.1 Procedure for the Identification of Bibliographic Information

To identify intralingual bibliographic information, traditional systems extract a title, author(s) name(s), year of publication, and the number of pages from a list of references, and then compare each field. In our task, we also extracted the same fields from each bibliographic information source, and compared each field. Here, the fields “number of pages” and “publication year” were directly comparable, while a machine translation technique was required to compare the fields “title” and “author(s) name(s)”. We can use several resources, such as dictionaries, for Japanese morphological analysis and for kana-kanji conversion to translate an author’s name. However, the “title” field was difficult to translate using general machine translation systems, because in general, a title is a large noun phrase containing some technical terms. Therefore, we identified interlingual bibliographic information using the following four steps.

1. Analyse the structure of a title based on some cue phrases,
2. Extract a series of nouns as technical terms from the titles,
3. Translate extracted technical terms based on a statistical translator for technical terms [1],
4. Identify interlingual bibliographic information based on the structure of the titles and the results of the translation of technical terms.

In the next section, we will elaborate on the first step.

2.2 Analysing the Structure of Titles

We analysed the structure of titles using cue phrases. The following text shows two examples of the results of our analysis of bibliographic information, shown in Section 1, using our method.

[Japanese title]
<METHOD> サポートベクトルマシン </METHOD> を用いた <HEAD> 自動要約 </HEAD>
[English title]
<HEAD>Automatic Summarization</HEAD> based on <METHOD>Support Vector Machines</METHOD>

In the Japanese title, the “METHOD” tag is assigned to “サポートベクトルマシン” (Support Vector Machines), because the cue phrase, “を用いた” (based on) appears just after it. In the English title, the “METHOD” tag is assigned to “Support Vector Machines,” because the cue phrase “based on” appears just before it. The “HEAD” tag is assigned to the last noun phrase in the Japanese

title and to the first noun phrase in the English title. Finally, we translated each tagged technical term using a statistical translator [1], and then compared them for each tag. We prepared 31 cue phrases in English and 165 cue phrases in Japanese to analyse the structure of the titles. Using these cue phrases, we manually made rules to assign 10 types of tag to each word in a title. We show a part of these tags and cue phrases in Table 1.

Table 1. A part of tags and cue phrases

Tag	Cue phrases (English)	Cue phrases (Japanese)
METHOD	based on, using, by	を用いた, に基づいた, による
RESTRICT	in, on, of	における, に関する, の
GOAL	towards, for	に向けて, のための
CONJ	and, or	と, や, 及び

3 Experiments

We conducted the following experiment to confirm the effectiveness of our method.

3.1 Data

We used the data set used for the Cross-lingual Information Retrieval (CLIR) Tasks in the first and second NTCIR Workshops (NTCIR-1 and NTCIR-2). This data set contains about 330,000 bibliographic information items, and XML-style tags were assigned to the title, author(s) name(s), year of publication, and abstract, all of which were written in Japanese and English. We randomly selected 750 pairs and used them for our task.

3.2 Alternatives

We identified bibliographic information using the following four methods.

- **Method 1.** Use a translation of technical terms in the titles.
- **Method 2.** Use Method 1 and a translation of the author(s) name(s).
- **Method 3.** Use Method 2 and a publication year.
- **Method 4.** Use Method 3 and any tag information.

As a baseline method, we identified bibliographic information using a translation of the author(s) name(s) and year of publication. To translate the author(s) name(s), we used ipadic, which contains 33,000 Japanese personal names and their pronunciations.

3.3 Evaluation

We evaluated the four methods and the baseline method using precision and recall, as defined in the following equations.

$$Precision = \frac{The\ number\ of\ titles\ that\ the\ system\ could\ detect\ correctly}{The\ number\ of\ titles\ that\ the\ system\ detected} \quad (1)$$

$$Recall = \frac{The\ number\ of\ titles\ that\ the\ system\ could\ detect\ correctly}{The\ number\ of\ titles\ that\ should\ be\ detected} \quad (2)$$

Table 2. A comparison of our methods and the baseline method.

	Precision (%)	Recall (%)
Baseline method	1.7	93.3
Method 1	3.4	29.4
Method 2	17.3	29.4
Method 3	91.9	29.4
Method 4	94.8	25.6

3.4 Results and Discussion

Table 2 shows our experimental results.

As can be seen from the Table 2, both "translation of author(s) name(s)" (Method 2) and "year of publication" (Method 3) are useful for improving the precision. The "tag information" (Method 4) was also useful, because the tag information improved the value of the precision by 2.9%. However, the tag information decreased the recall value by 3.8%.

The main reason for the low recall using our methods is that English titles are not always correct translations of Japanese titles. We show a typical example of a pair of English and Japanese titles that our method could not identify.

[**English**] Study of the dialogue model for an intelligent support system of group learning
[**Japanese**] 知的グループ学習支援システム のための' 対話モデルの研究

From the English title, our system extracted three terms "dialogue model", "intelligent support system", and "group learning", and correctly translated these into "対話モデル (dialogue model)", "知的学習支援システム (intelligent support system)", and "グループ学習 (group learning)", respectively. However, two terms, "intelligent support system" and "group learning", in the English title were integrated into a single noun phrase, "知的グループ学習支援システム (intelligent group learning support system)" in the Japanese title. These cases caused the low recall value. Analysing the structure of each term in the titles is required to resolve this problem. This study aims to identify elementary terms in a technical term, and to analyse the dependency relationship between them, and although there have been many studies in the fields of terminology study or natural language processing, research in this area is still very difficult.

4 Conclusions

We have proposed a method to identify interlingual bibliographic information based on analysing the structure of titles. In experiments, our method obtained a precision of 94.8% and a recall value of 25.6%, confirming the effectiveness of our method.

References

1. Fujii, A., and Ishikawa, T. Cross-Language Information Retrieval for Technical Documents. In Proceedings of the Joint ACL SIGDAT Conference on EMNLP and VLC, 29-37, (1999)
2. Nanba, H., Abekawa, T., Okumura, M., and Saito, S. Bilingual PRESRI: Integration of Multiple Research Paper Databases. In Proceedings of RIAO 2004, 195-211 (2004)